Turning Typological Universals into Annotation Auditors

Antoni Brosa-Rodríguez
Universitat Rovira i Virgili / Avda. Catalunya 35, 43002 Tarragona (Spain)
antoni.brosa@urv.cat

Relevant UniDive working groups: WG1

1 Introduction

Large-scale treebank collections such as Universal Dependencies (Nivre et al., 2020) have become central infrastructure for cross-linguistic research, enabling corpus-based validation of typological universals across many languages simultaneously (Ponti et al., 2019; O'Horan et al., 2016). The prevailing methodological direction uses corpora as evidence for or against typological claims (Gerdes et al., 2021; Levshina, 2019). In this work, we propose and demonstrate the inverse strategy: using well-established typological universals as diagnostic tools to audit and evaluate annotation quality in UD treebanks.

Morphological features represent a natural starting point for this approach. Unlike syntactic phenomena, morphological categories such as gender, number, tense, or person are directly encoded as discrete feature-value pairs in the CoNLL-U format, making them straightforwardly queryable and comparable across treebanks. This tractability, combined with the strong cross-linguistic support that many morphological universals enjoy (Corbett, 1991, 2000), makes them ideal candidates for large-scale annotation auditing—more immediate and reliable than syntax-based universals, where structural variation and annotation conventions interact in more complex ways.

We ground this approach in four of Greenberg's (1963) morphological universals, selected for their direct encoding in UD morphological features and their strong typological backing:

- **Universal 30**: If the verb has categories of person-number or gender, it always has tense-mood categories.
- **Universal 31**: If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.
- **Universal 42**: All languages have pronominal categories involving at least three persons and two numbers.
- **Universal 45**: If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular also.

Each universal encodes an implication that functions as a falsifiable prediction over UD annotations: any treebank violating the predicted pattern constitutes a flagged case, to be examined as either a genuine typological exception or an annotation error. This allows the universal to lose its axiomatic weight and simply serve as a warning once the annotation has been done.

2 Methodology

We formalize each universal as morphosyntactic queries over UD feature inventories, targeting combinations of GENDER, NUMBER, PERSON, TENSE, and MOOD across relevant UPOS categories. Queries were designed following the formal query language of Grew-Match (Guillaume, 2021) and executed programmatically over all available treebanks in UD v2.15 and v2.17 (up to 339 treebanks, 186 languages). For each treebank and universal, we record compliance or violation and quantify occurrence frequencies. The linguistic sample covers all macro-areas, though with known genealogical biases: Indo-European languages are overrepresented, while Austronesian, Papuan, and North American families have minimal coverage (Miestamo et al., 2016). Compared to WALS (Dryer and Haspelmath, 2013), the distribution of dominant word order types is reasonably balanced, with a slight overrepresentation of SVO.

Annotation errors are classified into two types: overgeneration errors, where features are incorrectly assigned to elements that should not bear them (e.g., through automatic conversion artifacts or overgeneralization from specific contexts), and data omission errors, where features are inconsistently or incompletely annotated where they should be present (e.g., due to annotation fatigue or disagreement over implicit vs. explicit marking). Both

types distort typological analyses in opposite directions: overgeneration creates false evidence supporting a universal or spurious violations of another, while omission obscures genuine linguistic patterns.

3 Results and Discussion

Universal 45 has been studied extensively from grammar-based evidence (Plank and Schellinger, 1997; Siewierska, 2013; Berg, 2024; Plank, 2025). Corpus-based verification confirms the implication: of the 139 treebanks with plural gender distinctions in pronouns, 137 (98.6%) also exhibit them in the singular. The two exceptions (Hausa & Maghrebi Arabic French) involve few occurrences, strongly suggesting annotation errors. Extending the analysis across 13 UPOS categories shows that invariable categories (ADP, ADV, CCONJ, SCONJ, INTJ), expected to show 0 occurrencies, yield the highest violation rates—up to 42.9% for SCONJ—since no language is expected to mark gender on these PoS. Recurring error patterns include prepositional contractions inheriting article features in some treebanks, while in other treebanks from the same language don't (e.g., Italian *della* ['of the'] annotated as ADP with GENDER=FEM), homophone merging between conjunctions and copular verbs, and erroneous feature propagation. Even within inflecting categories, violations often trace to annotation conventions rather than linguistic reality.

Regarding U31, Spanish treebanks show a systematic overgeneration: perfect participles in compound verb forms are incorrectly annotated with GENDER=MASC across multiple treebanks, conflating Spanish with languages that genuinely mark gender on verbs. While it may be understood that the annotation should be made individually in the specific form rather than the compound form, this is not always the case in languages comparable to Spanish. This creates a discrepancy that sets such languages apart. Conversely, invariant adjectives like *oficial* receive inconsistent gender annotation both within Spanish treebanks and across related Romance languages, reflecting unresolved tension in UD guidelines between marking only morphologically overt features versus syntactically inferable ones (Haspelmath, 2010).

For Universal 42, Dutch treebanks (Alpino and LassySmall) systematically omit NUMBER features on all personal pronouns, which could lead to misclassifying Dutch as lacking pronominal number distinctions. For Universal 30, Spanish treebanks show undergeneration in tense-mood-aspect attribution co-occurring with the gender overgeneration above, producing a distorted typological profile simultaneously in both directions.

4 The Inverted Paradigm and Its Advantages

The standard pipeline in computational typology flows from corpora to typological claims (Gerdes et al., 2021; Liu, 2010; Guzmán-Naranjo and Becker, 2018). We argue for a complementary, bidirectional pipeline in which typological universals are deployed as annotation quality metrics. Morphological universals are the natural entry point: their feature-level encoding in UD makes them directly queryable, their violation conditions are precisely defined, and their cross-linguistic robustness provides a reliable baseline against which anomalies can be measured (Corbett, 1991, 2000). A single formalized universal applied computationally across all treebanks surfaces candidate errors at scale, provides language-agnostic diagnostics, and simultaneously yields quantitative cross-linguistic data unavailable from grammar-based typology (Dryer and Haspelmath, 2013). The methodology is fully replicable and extensible to other universals encodable through UD morphosyntactic features (Spencer, 2002). Beyond typology, annotation inconsistencies may distort any other linguistic metric.

5 Conclusions

We present a methodological inversion with practical consequences for corpus annotation: rather than asking what treebanks tell us about linguistic universals, we ask what linguistic universals tell us about treebanks. Applied to four of Greenberg's morphological universals across UD, this approach confirms typological robustness while exposing systematic annotation problems—both overgeneration and omission—with documented effects on typological validity. Morphological features, as the most tractable and precisely encoded dimension of UD annotation, represent the ideal first step toward a broader program of typology-driven annotation auditing. It remains to be checked (as a topic for future research) whether this methodology can be applied to more complex cases involving syntax and its development using Surface Universal De-

pendencies (SUD). This review we conduct allows us to continue improving the annotation of the corpora and to highlight some of the issues that may underlie exceptions to linguistic universals or may be giving us a distorted picture of a language if we are unaware of its true characteristics.

References

Thomas Berg. 2024. Gender, number and person: A three-way interaction. *Linguistic Typology*, 28(3):537–565.

Greville Corbett. 1991. *Gender*. Cambridge University Press.

Greville Corbett. 2000. *Number*. Cambridge University Press.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa*, 6(1).

Joseph Greenberg. 1963. *Universals of Language*. The M.I.T. Press, Cambridge.

Bruno Guillaume. 2021. Graph matching and graph rewriting: Grew tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175.

Matías Guzmán-Naranjo and Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories*.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86:663–687.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23:533–572.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Matti Miestamo, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 4034–4043.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016*, pages 1297–1308.

Frans Plank. 2025. Typology re-done: Greenberg nos. 37, 45, 34. *Linguistic Typology*. Advance online publication.

Frans Plank and Wolfgang Schellinger. 1997. The uneven distribution of genders over numbers: Greenberg nos. 37 and 45. *Linguistic Typology*, 1:53–101.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45:1–56.

Anna Siewierska. 2013. Gender distinctions in independent personal pronouns. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Andrew Spencer. 2002. Gender as an inflectional category. *Linguistics*, 38:279–312.